BERTScore

Paper analysis by Satya Swaroop Gudipudi (2022900024)

ICLR 2020

N-gram matching approaches:

Formally, let S n x and S n xˆ be the lists of token n-grams (n ∈ Z+) in the reference x and candidate xˆ sentences. The number of matched n-grams is :

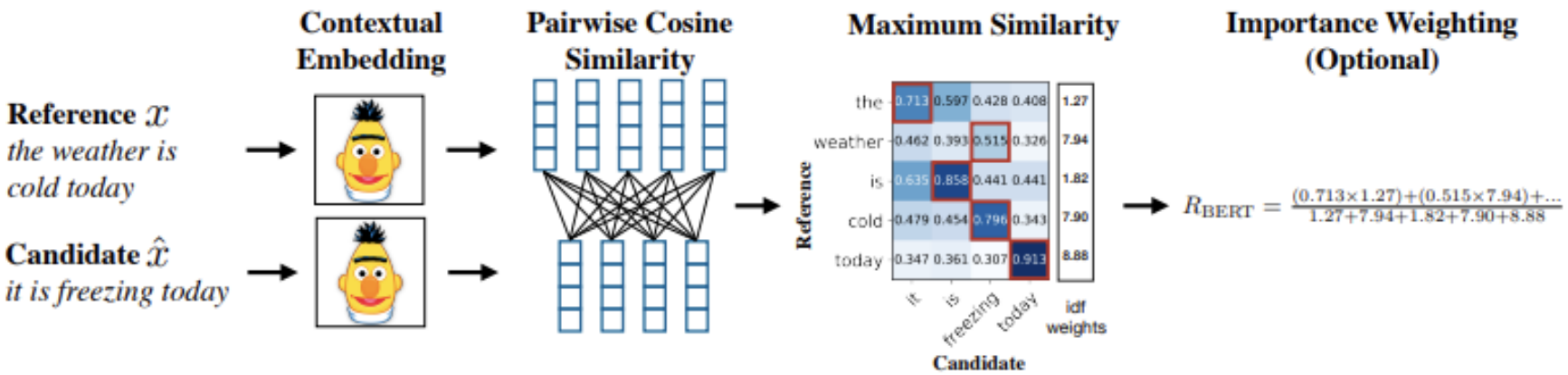
$$\sum_{w \in S^n_{\hat{x}}} \mathbb{I}[w \in S^n_x],$$

$$\text{Exact-P}_n = \frac{\sum_{w \in S^n_{\hat{x}}} \mathbb{I}[w \in S^n_x]}{|S^n_{\hat{x}}|} \quad \text{and} \quad \text{Exact-R}_n = \frac{\sum_{w \in S^n_x} \mathbb{I}[w \in S^n_{\hat{x}}]}{|S^n_x|}.$$

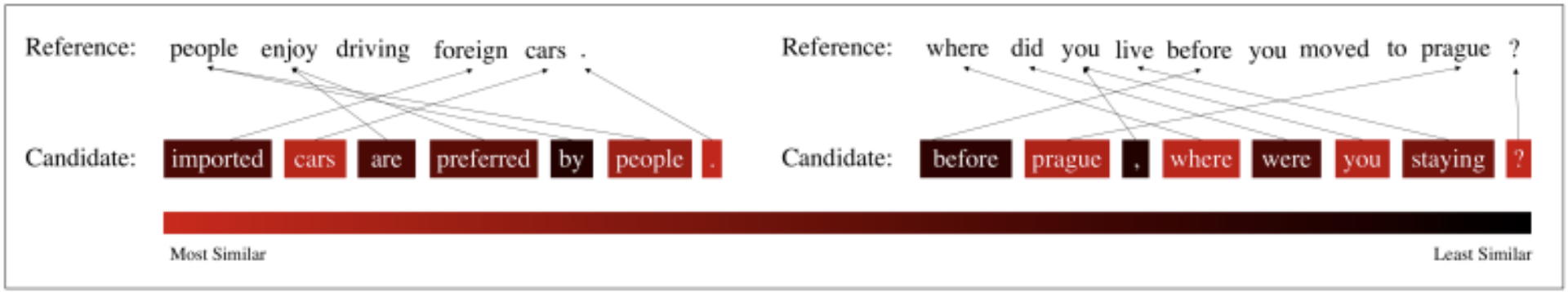
Popular metrics:

- BLUE
- METEOR
- ROUGE
- ΔBLEU
- WMD based on Earth Moving Distance

BERTScore



Pairwise interactions example:



Pros

- Able to take semantics into account while evaluating the candidates with references
- It is giving weighted evaluation to weigh important tokens that are rare using IDF
- It is differentiable unlike bleu hence it can be part of training process as optimization

Cons

- Finds difficulty to detect factual errors
- Fails to identify 5 feet 11 inches as equivalent with 1.80 meters
- Less guidelines on when to use what type of configuration of BERTScore for evaluating models.
- The evaluation would be limited by the underlying base model limitations example BERT is not good to handle negation like “Not” and longer text.

Improvements

- It would be insightful to evaluate the metric on more diverse dataset, for example, It is not clear how BERTScore could handle domain specific or out of vocab or spelling mistake (perturbed text).
- Integration of external knowledge to provide the model with additional and latest information that can help in factual evaluation as well.
- As BERT is an encoder model and BERTScore is evaluating on generated text. It will be interesting to compare BERTScore with LLMs evaluation that is evaluating the sentence pairs with help of LLMs.
- If BERTScore can be improved to evaluate the overall meaningful sense and consistency of the generated sentence as an extension to evaluating in absence of reference sentences at inference time.